Mapping ICD-10 and ICD-10-CM Codes to Phecodes:

Original Paper

Patrick Wu, BS[1,5,a], Aliya Gifford, PhD[1,a], Xiangrui Meng, PhD[2,a], Xue Li, PhD[2], Harry Campbell, MD[2], Tim Varley, BSc[3], Juan Zhao, PhD[1], Robert Carroll, PhD[1], Lisa Bastarache, MS[1], Joshua C Denny, MD MS[1,6], Evropi Theodoratou, PhD[2,4], Wei-Qi Wei, MD PhD[1]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA
[2]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK
[3]Public Health and Intelligence SBU, National Services Scotland, Edinburgh, UK
[4]Edinburgh Cancer Research Centre, Institute of Genetics and Molecular Medicine, The University of Edinburgh, Edinburgh, UK
[5]Medical Scientist Training Program, Vanderbilt University School of Medicine, Nashville, TN, USA
[6]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[a]Authors made equal contributions to the paper

Correspondence to:
Wei-Qi Wei, MD PhD
Email: wei-qi.wei@vumc.org
Department of Biomedical Informatics
Vanderbilt University Medical Center
2525 West End Ave., Suite 1500
Nashville, TN 37203
Tel: (615) 343-1956
Fax: (615) 936-0102

Developing and Evaluating Mappings of ICD-10 and ICD-10-CM Codes to Phecodes

Abstract

Background:
The phecode system was built upon the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) for phenome-wide association studies (PheWAS) in the electronic health record (EHR).

Objectives:
We present our work on the development and evaluation of maps from ICD-10 and ICD-10-CM codes to phecodes.

Methods:
We mapped ICD-10 and ICD-10-CM codes to phecodes using a number of methods and resources, such as concept relationships and explicit mappings from the Centers for Medicare & Medicaid Services, the Unified Medical Language System, Observational Health Data Sciences and Informatics, Systematized Nomenclature of Medicine - Clinical Terms, and the National Library of Medicine. We assessed the

coverage of the maps in two databases: Vanderbilt University Medical Center (VUMC) using ICD-10-CM and the UK Biobank (UKBB) using ICD-10. We assessed the fidelity of the ICD-10-CM map in comparison to the gold-standard ICD-9-CM phecode map by investigating phenotype reproducibility and conducting a PheWAS.
Results:
We mapped >75% of ICD-10 and ICD-10-CM codes to phecodes. Of the unique codes observed in the UKBB (ICD-10) and VUMC (ICD-10-CM) cohorts, >90% were mapped to phecodes. We observed 70-75% reproducibility for chronic diseases and <10% for an acute disease for phenotypes sourced from the ICD-10-CM phecode map. Using the ICD-9-CM and ICD-10-CM maps, we conducted a PheWAS with a lipoprotein(a) (*LPA*) genetic variant, rs10455872, which replicated two known genotype-phenotype associations with similar effect sizes: coronary atherosclerosis (ICD-9-CM: *P*=1.96E-15, odds ratio (OR) = 1.60, 95% confidence interval (CI): 1.43-1.80 vs. ICD-10-CM: *P*=8.63E-16, OR = 1.60, 95% CI: 1.43-1.80) and chronic ischemic heart disease (ICD-9-CM: *P*=4.18E-10, OR = 1.56, 95% CI: 1.35-1.79 vs. ICD-10-CM: *P*=5.21E-05, OR = 1.47, 95% CI: 1.22-1.77).
Conclusions:
This study introduces the “beta” versions of ICD-10 and ICD-10-CM to phecode maps that enable researchers to leverage accumulated ICD-10 and ICD-10-CM data for PheWAS in the EHR. The maps are available from https://phewascatalog.org and incorporated in the PheWAS R package, https://github.com/PheWAS/PheWAS.

Keywords:
electronic health record; genome-wide association study; phenome-wide association study; phenotyping

Introduction

Background

Electronic health records (EHRs) have become a powerful resource for biomedical research in the last decade, and many studies based on EHR data have used International Classification of Diseases (ICD) codes [1]. When linked to DNA biobanks, healthcare information in EHRs is a tool to discover genetic associations using billing codes in phenotyping algorithms. The phenome-wide association study (PheWAS) paradigm was introduced in 2010 as an approach that scans across a range of phenotypes, similar to genome-wide association studies. Studies using PheWAS have replicated hundreds of known genotype-phenotype associations and discovered dozens of new ones [2–12]. The initial version of phecodes consisted of 733 custom groups of ICD Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis codes. The most recent iteration of phecodes consists of 1,866 hierarchical phenotype codes that map to 15,558 ICD-9-CM codes [13,14]. However, many health systems and international groups use ICD-10 or ICD-10-CM codes [15], therefore necessitating a new phecode map.

Transition from ICD-9 to ICD-10

In 1979, the World Health Organization (WHO) developed ICD-9 to track mortality and morbidity. To improve its application to clinical billing, the United States National Center

for Health Statistics (NCHS) modified ICD-9 codes to create ICD-9-CM, whose end-of-life date was scheduled around the year 2000, but was delayed until October 2015 [15]. In 1990, the WHO developed ICD-10 [16], which the NCHS used to create ICD-10-CM to replace ICD-9-CM.

Moving from ICD-9-CM to ICD-10-CM led to major structural changes in the coding system. First, the structure moved from a broadly numeric-based system in ICD-9-CM (e.g. 474.11 for "Hypertrophy of tonsils alone") to an alphanumeric system in ICD-10-CM (e.g. J35.1 for the same condition). Second, ICD-10-CM contains much more granular information than ICD-9-CM, as seen with the approximately tenfold increase in the number of diabetes-related codes in ICD-10-CM. ICD-10-CM also differs from ICD-9-CM in terms of semantics and organization [15,17].

Compared to ICD-10, ICD-10-CM has more codes and granularity. While the 2018AA Unified Medical Language System (UMLS) [18] contains 94,201 unique ICD-10-CM codes, it has 12,027 unique ICD-10 codes after exclusion of range codes (e.g. ICD-10-CM A00-A09). Further, there are ICD-10 codes that do not exist in ICD-10-CM, and vice versa, such as ICD-10 A16.9 "Respiratory tuberculosis unspecified, without mention of bacteriological or histological confirmation", which has no ICD-10-CM equivalent.

Prior Work

To develop the original phecode system, one or more related ICD-9-CM codes were combined into distinct diseases or traits. For example, three depression-related ICD-9-CM codes 311, 296.31, and 296.2 are condensed to phecode 296.2 "Depression". With the help of clinical experts in disparate domains, such as cardiology and oncology, we have iteratively updated the phecode groupings [19].

The phecode scheme is unique because it has built-in exclusion criteria to prevent contamination by cases in the control cohort. This is an important feature, as case contamination of control groups decreases the statistical power to find genotype-phenotype associations [20]. For each disease phenotype, we defined exclusion criteria by using our clinical knowledge and by consulting physician specialists.

An example for how users can use phecode exclusion criteria is illustrated by a type 2 diabetes study in the EHR. To define cases of type 2 diabetes, users include patients with ICD codes that map to phecode 250.2 "Type 2 diabetes". To create the control cohort, they include patients without phenotypes in the "DIABETES" group, which is comprised of phecodes in the range of 249-250.99. This prevents contamination of the control group by patients with diseases such as "Type 1 diabetes" (phecode 250.1) and "Secondary diabetes mellitus" (phecode 249). Excluded patients also include those with signs and symptoms commonly associated with type 2 diabetes, such as "Abnormal glucose" (phecode 250.4), which may indicate someone who has not yet been diagnosed with diabetes.

Though the phecode system is effective at replicating and identifying novel genotype-phenotype associations, PheWAS have largely been limited to using ICD-9-CM codes.

A few studies have mapped ICD-10 codes to phecodes by converting ICD-10 to ICD-9-CM, and then mapping the converted ICD-9-CM codes to phecodes [3,10]. However, these studies limited their mappings to ICD-10 (non-CM) codes, did not provide a map to translate ICD-10-CM codes to phecodes, and did not evaluate the accuracy of these maps.

Goal of this Study

In this study, we developed and evaluated maps of ICD-10 and ICD-10-CM codes to phecodes. The primary aims of this study were to create an initial "beta" map to perform PheWAS using ICD-10 and ICD-10-CM codes and to focus the analyses on PheWAS-relevant codes. Our goal was to demonstrate that researchers should expect similar results from the ICD-10-CM phecode map compared to the gold-standard ICD-9-CM map. To accomplish this goal, we investigated phecode coverage, phenotype reproducibility, and the results from a PheWAS.

Methods

Databases

In this study, we used data obtained from the Vanderbilt University Medical Center (VUMC) and UK Biobank (UKBB) [25] databases. The VUMC EHR contains clinical information derived from the medical records of >3 million unique individuals. The UKBB is a prospective longitudinal cohort study designed to investigate the genetic and environmental determinants of diseases in UK adults. Between 2006-2010, the study recruited >500,000 men and women aged 40-69 years. Participants consented to allow their data to be linked to their medical records. EHR records of UKBB were obtained under an approved data request application (ID:10775).

At the time of this study, VUMC had >2.5 years of ICD-10-CM data (~2015-10-01 to 2017-06-01), while the UKBB had >2 decades of ICD-10 data [21] (~1995-04-01 to 2015-03-31). VUMC includes codes for inpatient and outpatient encounters, whereas UKBB codes in this study are only inpatient codes.

Mapping ICD-10-CM and ICD-10 Codes to Phecodes

We extracted ICD-10-CM codes from the 2018AA release of the UMLS [18], and used a number of automated methods to translate ICD-10-CM diagnosis codes to phecodes (Figure 1). We mapped 515 ICD-10-CM codes directly to phecodes by matching code descriptions regardless of capitalization, e.g. ICD-10-CM H52.4 "Presbyopia" to phecode 367.4 "Presbyopia". We mapped 82,287 ICD-10-CM codes indirectly to phecodes using the existing ICD-9-CM phecode map [14]. To convert ICD-10-CM codes indirectly to phecodes, we used General Equivalence Mappings (GEMS) provided by the Centers for Medicare & Medicaid Services, that maps ICD-10-CM to ICD-9-CM and vice versa [22]. We included both equivalent and non-equivalent GEMS mappings (i.e. where the "approximate" flag was either "0" or "1"). As an example of this indirect approach, to map ICD-10-CM E11.9 "Type 2 diabetes mellitus without complications" to phecode 250.2 "Type 2 diabetes": ICD-10-CM E11.9 to ICD-9-CM 250.0 "Diabetes mellitus without mention of complication" to phecode 250.2.

Since the GEMS do not provide mappings for all ICD-10-CM codes [17], we complemented this approach with UMLS semantic mapping [23], Observational Health Data Sciences and Informatics (OHDSI) concept relationships [24,25], and National Library of Medicine (NLM) maps [26]. In this approach to indirect mapping, we first mapped ICD-10-CM codes to Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) through UMLS Concept (CUI) equivalents, which were then converted to ICD-9-CM through either UMLS CUI equivalents [18,23], OHDSI [24], or NLM maps [26]. For example, ICD-10-CM L01.00 “Impetigo, unspecified” to CUI C0021099 to SNOMED CT 48277006 to OHDSI Concept ID 140480 to OHDSI Concept ID 44832600 to ICD-9-CM 684 to phecode 686.2 “Impetigo”.

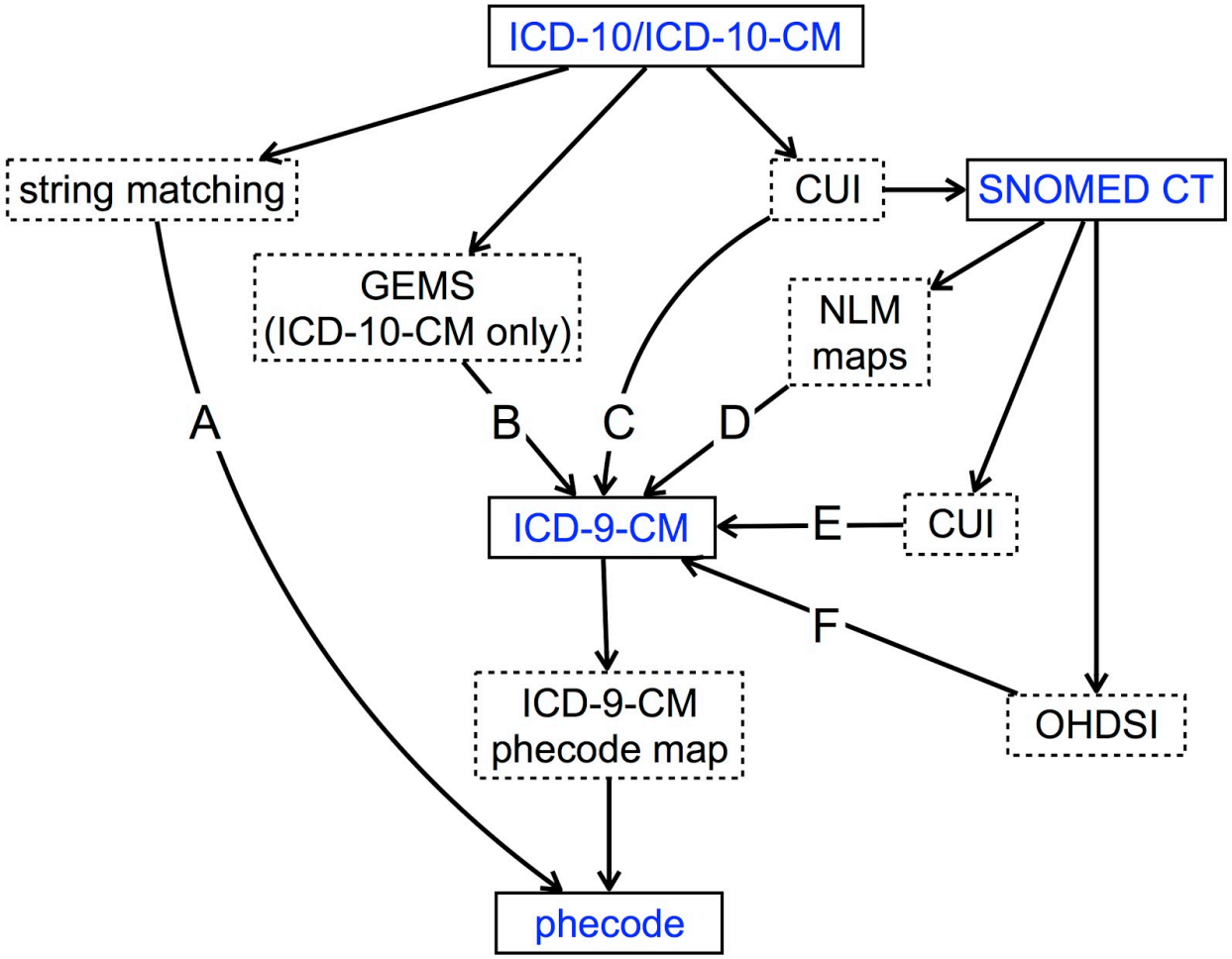


Figure 1. Mapping strategy for ICD-10 (non-CM) and ICD-10-CM diagnosis codes to phecodes. We mapped ICD-10-CM codes directly by matching code descriptions (path A) or indirectly to phecodes, using a number of manually-validated mapping resources (paths B, C, D, E, and F). In path D, we used NLM’s SNOMED CT to ICD-9-CM one-to-one and many-to-one maps [26]. To map ICD-9-CM codes to phecodes, we applied Phecode Map 1.2 with ICD-9 Codes (ICD-9-CM phecode map) [14]. Boxes with solid lines indicate clinical terminologies, and those with dashed lines describe the resources

and mapping methods used. ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification. SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms. GEMS: General Equivalence Mappings. UMLS: Unified Medical Language System. OHDSI: Observational Health Data Sciences and Informatics. CUI: Concept Unique Identifier. NLM: National Library of Medicine.

There were two general instances when an ICD-10-CM code mapped to more than one phecode. First, some ICD-10-CM codes mapped to a parent phecode and one if its child phecodes that was lower in the hierarchy. To maintain the granular meanings of ICD-10-CM codes, we only kept the mappings to child phecodes, a decision that we could make due to the hierarchical structure of phecodes. For example, ICD-10-CM I10 "Essential (primary) hypertension" was mapped to phecodes 401 "Hypertension" and 401.1 "Essential hypertension", but we only kept the mapping to phecode 401.1. Second, we kept all the mappings for ICD-10-CM codes that were translated to phecodes that were not in the same family. This can be seen in the mapping of ICD-10-CM D57.812 "Other sickle-cell disorders with splenic sequestration" to phecodes 282.5 "Sickle cell anemia" and 289.5 "Diseases of spleen". This latter association created a polyhierarchical nature to phecodes that did not previously exist.

To map ICD-10 (non-CM) codes to phecodes, we used ICD-10 codes also from the 2018AA UMLS [18]. ICD-10 codes were mapped to phecodes in a similar manner to ICD-10-CM, but since a GEMS to translate ICD-10 to ICD-9-CM was not available, we used only string matching and previously manually-reviewed resources from the UMLS [23], NLM [26], and OHDSI [24,25].

Evaluation of Phecode Coverage of ICD-10 and ICD-10-CM in UKBB and VUMC
To evaluate the phecode coverage of ICD-10 and ICD-10-CM source codes in UKBB and VUMC, respectively, we calculated the number of source codes in the 2018AA UMLS, number of source codes mapped to phecodes, and number of mapped and unmapped source codes that were used in the two EHRs (Figure 2). To identify potential limitations of our automated mapping approach, two authors with clinical training (P.W., W.Q.W.) manually reviewed all the unmapped ICD-10 and ICD-10-CM codes that were used at UKBB and VUMC, respectively.

Comparison of Phenotypes Generated from the ICD-10-CM Phecode Map
We aimed to provide evidence that the ICD-10-CM phecode map resulted in phenotypes similar to those sourced from the ICD-9-CM phecode map. First, we selected 357,728 patients in the VUMC EHR who had ≥1 ICD-9-CM and ≥1 ICD-10-CM

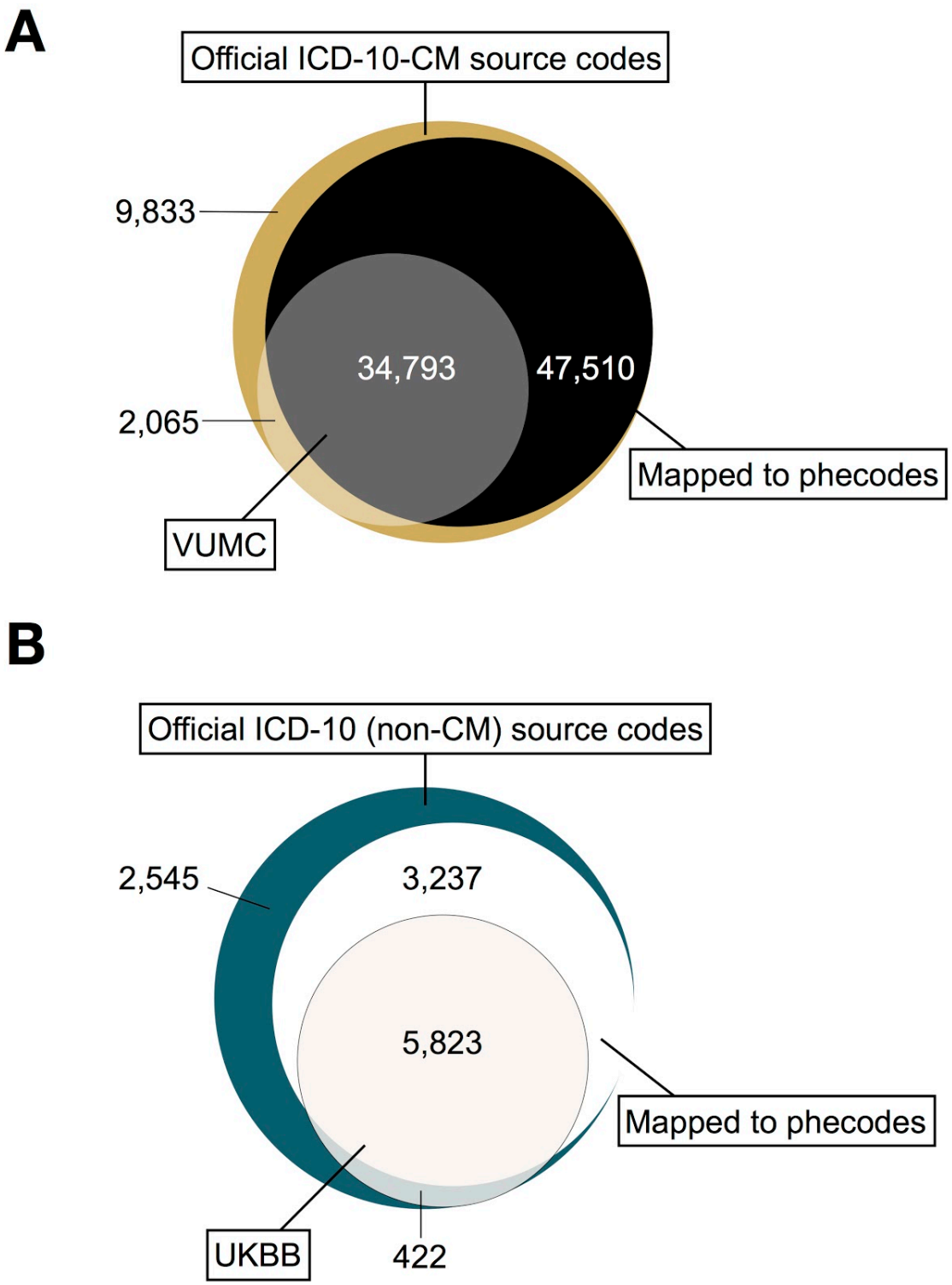


Figure 2. Counts of distinct ICD-10-CM source codes at VUMC and ICD-10 (non-CM) source codes in UKBB. (A) Number of unique ICD-10-CM codes in each category. For example, there were 34,793 unique codes (grey section) that were in the official ICD-10-CM system, observed in the VUMC dataset, and mapped to phecodes. (B) Number of unique ICD-10 codes in each category. For example, there were 5,823 unique codes (off-white section) that were in the official ICD-10 system, observed in the UKBB dataset,

and mapped to phecodes. VUMC: Vanderbilt University Medical Center. UKBB: UK Biobank.

codes in two 18-month windows. We selected windows to occur prior to and after VUMC's transition to ICD-10-CM. To reduce potential confounders, we left a six-month buffer after ICD-9-CM was replaced with ICD-10-CM. Further, the ICD-10-CM observation window ended before VUMC switched from its locally developed EHR [27] to the Epic system. This created two windows ranging from 2014-01-01 to 2015-06-30 for ICD-9-CM, and 2016-01-01 to 2017-06-30 for ICD-10-CM (Figure 3). The final cohort consisted of 55.10% female with mean (standard deviation, SD) 45 (25) years old. From the two observation periods, we extracted all ICD-9-CM and ICD-10-CM codes for each patient. We then mapped these codes to phecodes using the ICD-9-CM phecode [14] and ICD-10-CM phecode maps.

Figure 3. Timeline of the two 18-month periods from which ICD-9-CM and ICD-10-CM codes from VUMC were analyzed. The cohort of 357,728 patients had at least one ICD-9-CM and one ICD-10-CM code in the respective 18-month windows.

We used the patient cohort to test our hypothesis that the ICD-10-CM phecode map created phenotype definitions that were comparable to those generated using the gold-standard ICD-9-CM phecode map. For this analysis, we used four common chronic diseases (Hypertension, Hyperlipidemia, Type 1 Diabetes, and Type 2 Diabetes) and chose one acute disease (Intestinal infection) as a negative control. We expected that a large majority of the chronic disease patients and small minority of the acute disease patients from the ICD-9-CM era would reproduce the same phenotypes during the ICD-10-CM era. We defined the phenotype cases as follows: Hypertension with phecodes 401.* ("*" means one or more digits or a period); Hyperlipidemia, phecodes 272.*; Type 1 diabetes, phecodes 250.1*; Type 2 diabetes, phecodes 250.2*; Intestinal infection, phecodes 008.*.

For each phenotype, we reported the number of ICD-9-CM cases and the number of those individuals who were also ICD-10-CM cases. To identify the possible reasons for individuals who were not identified as phenotype cases in the ICD-10-CM period, two authors with clinical training (P.W., W.Q.W.) manually reviewed the EHRs of ten randomly selected patients from each chronic disease group, except Type 1 diabetes, for a total of thirty patients.

Comparative PheWAS Analysis of Lipoprotein(a) (*LPA*) Single-nucleotide polymorphism (SNP)

To evaluate the accuracy of the ICD-10-CM phecode map, we performed two PheWAS on an *LPA* genetic variant (rs10455872) using mapped phecodes from ICD-9-CM and ICD-10-CM. The *LPA* SNP is associated with increased risks of developing hyperlipidemia and cardiovascular diseases [28–30].

We used data from BioVU, the de-identified DNA biobank at VUMC to conduct the PheWAS [31]. We identified 13,900 adults (56.90 % female, 59 (15) years old in 2014), who had rs10455872 genotyped, and at least one ICD-9-CM and ICD-10-CM code in their respective time windows. For rs10455872, we observed 86.7% AA, 12.8% AG, and 0.5% GG. We used 1,632 phecodes that overlapped in the time windows for PheWAS using the R PheWAS package [13] with binary logistic regression, adjusting for age, sex, and race.

Results

Phecode Coverage of ICD-10-CM and ICD-10 in VUMC and UKBB

Of all possible ICD-10-CM codes [32], 82,303 (87.37%) mapped to at least one phecode, with 7,881 (8.37%) mapping to >1 phecode. For example, ICD-10-CM I25.708 “Atherosclerosis of coronary artery bypass graft(s), unspecified, with other forms of angina pectoris” mapped to phecodes 411.3 “Angina pectoris” and 411.4 “Coronary atherosclerosis”. Of all possible ICD-10 codes, 9,060 (75.33%) mapped to at least one phecode, and 289 (2.40%) mapped to >1 phecode. For example, ICD-10 code B21.1 “HIV disease resulting in Burkitt lymphoma” maps to phecodes 071.1 “HIV infection, symptomatic” and 202.2 “Non-Hodgkins lymphoma”.

Among the 36,858 ICD-10-CM codes used at VUMC, 34,793 (94.40%) codes were mapped to phecodes. In the UKBB, 5,823 (93.24%) of the ICD-10 codes mapped to phecodes (Table 1, Figure 2). Considering all the instances of ICD-10-CM and ICD-10 codes used at each site, we generated a total count of unique codes grouped by patient and date, and those codes that mapped to phecodes (Table 1). Among the total number of codes used, 89.72% of ICD-10-CM and 83.68% of ICD-10 codes were mapped to phecodes.

Table 1. ICD-10-CM and ICD-10 codes data summary.

	ICD-10-CM (No.) (VUMC)	ICD-10 (No.) (UKBB)
Official classification systems		
Unique codes	94,201	12,027
Unique codes mapped	82,303 (87.37%)	9,060 (75.33%)

Official codes used in cohorts		
Unique codes	36,858	6,245
Unique codes mapped	34,793 (94.40%)	5,823 (93.24%)
Total patients (with ICD-10-CM or ICD-10 codes)	651,649	391,181
Total instances of all ICD codes	19,682,697	5,114,363
Instances mapped to phecodes	17,658,470 (89.72%)	4,279,544 (83.68%)

Analysis of Unmapped ICD-10 and ICD-10-CM Codes

Majority of the unmapped ICD-10 codes used in the UKBB dataset represented medical concepts related to personal (i.e. past medical history) or family history of disease. For ICD-10-CM, removing codes used at VUMC that we expected to be unmapped (i.e. local or supplementary classification codes) left 2,065 ICD-10-CM codes that did not map to a phecode. After excluding X, Y, and Z codes (1,395 codes), 670 codes remained, majority of which represented either “external causes of morbidity” or “factors influencing health status and contact with health services”. All of the remaining unmapped ICD-10-CM codes in this cohort had <200 unique individuals (i.e. <.1% of the cohort), and majority of the ICD-10-CM codes with >10 unique individuals were phenotypes that are most likely due to non-genetic factors. For example, 287 (59.2%) of the unmapped ICD-10-CM codes represented external causes of morbidity, such as assault and injuries due to motor vehicle accidents.

Reproducibility Analysis of the ICD-10-CM Phecode map

In the defined 18-month time windows, a cohort 357,728 patients had both ICD-9-CM and ICD-10-CM codes (Figure 3). For the chronic diseases, 70-75% of individuals with the relevant phecodes in the ICD-9-CM observation period also had the same phecodes of interest during the ICD-10-CM period. On the contrary, for the reproducibility analysis with an acute disease, we observed that <10% of individuals who had phecodes 008.* (Intestinal infection) in the ICD-9-CM period also had the same phecodes in the ICD-10-CM period (Table 2).

Table 2. ICD-10-CM phecode map reproducibility analysis.

Phenotype	Phecodes[a]	No. ICD-9-CM cases	No. Individuals (%), (ICD-10-CM case | ICD-9-CM case)[b]
Hypertension	401.*	65,216	49,468 (75.85%)

Hyperlipidemia	272.*	51,187	36,187 (70.70%)
Type 1 diabetes	250.1*	5,782	4,412 (76.31%)
Type 2 diabetes	250.2*	25,077	19,066 (76.03%)
Intestinal infection	008.*	3,410	273 (8.01%)

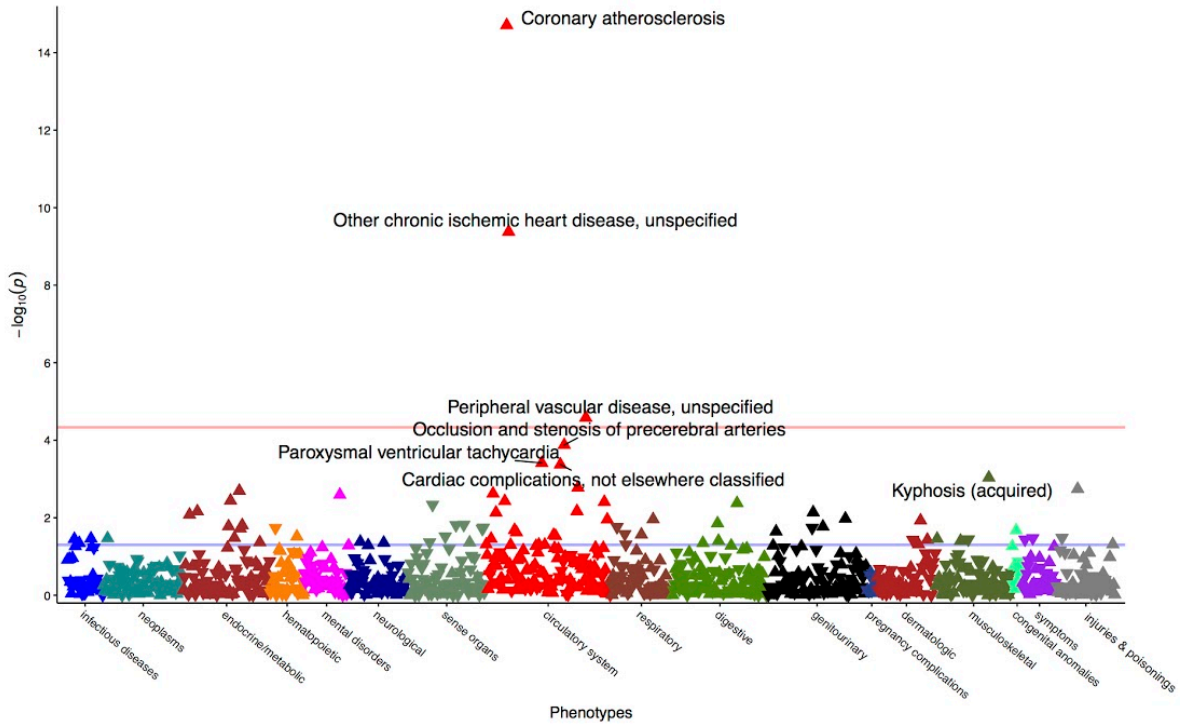
[a]In the phecode column, "*" means ≥1 digits or a period, e.g. phecode 401.* = phecodes 401, 401.1, 401.3, 401.22, 401.21, 401.2.
[b]In the last column, "ICD-10-CM case | ICD-9-CM case" indicates patients who were cases for the phenotype of interest during the ICD-9-CM period who were also ICD-10-CM cases.

To identify the reasons that may explain why some patients were not identified as cases for the phenotype of interest during the ICD-10-CM period, we manually reviewed their medical records. Thirty patients were selected for review, ten each from the Hypertension, Hyperlipidemia, and Type 2 diabetes cohorts (Multimedia Appendix 1). We found that none of the patients had a relevant ICD-10-CM code for the phenotype being studied in the 18-month observation period. Reasons for patients not being ICD-10-CM cases include: patients were labeled with the relevant ICD-10-CM code(s) outside of the short ICD-10-CM observation window (8 patients), patients had <2 visits at VUMC during the ICD-10-CM period and/or were only seen by physician specialists (10 patients; e.g. patient with hypertension was only seen by their neurologist during the ICD-10-CM period), and patients were inconsistently diagnosed (2 people; e.g. patient with Type 1 diabetes given Type 2 diabetes ICD-9-CM code). No cases were missed due to errors in the ICD-10-CM phecode map.

Comparative PheWAS Analysis of *LPA* SNP, rs10455872

To further evaluate the ICD-10-CM phecode map, we performed and compared the results of PheWAS analyses for rs10455872. One PheWAS was conducted using the ICD-9-CM map and another was conducted using the ICD-10-CM map. Both analyses replicated previous findings with similar effect sizes: coronary atherosclerosis (ICD-9-CM: *P*=1.96E-15, odds ratio (OR) = 1.60, 95% confidence interval (CI): 1.43-1.80 vs. ICD-10-CM: *P*=8.63E-16, OR = 1.60, 95% CI: 1.43-1.80) and chronic ischemic heart disease (ICD-9-CM: *P*=4.18E-10, OR = 1.56, 95% CI: 1.35-1.79 vs. ICD-10-CM: *P*=5.21E-05, OR = 1.47, 95% CI: 1.22-1.77) (Figure 4).

ICD-9-CM



ICD-10-CM

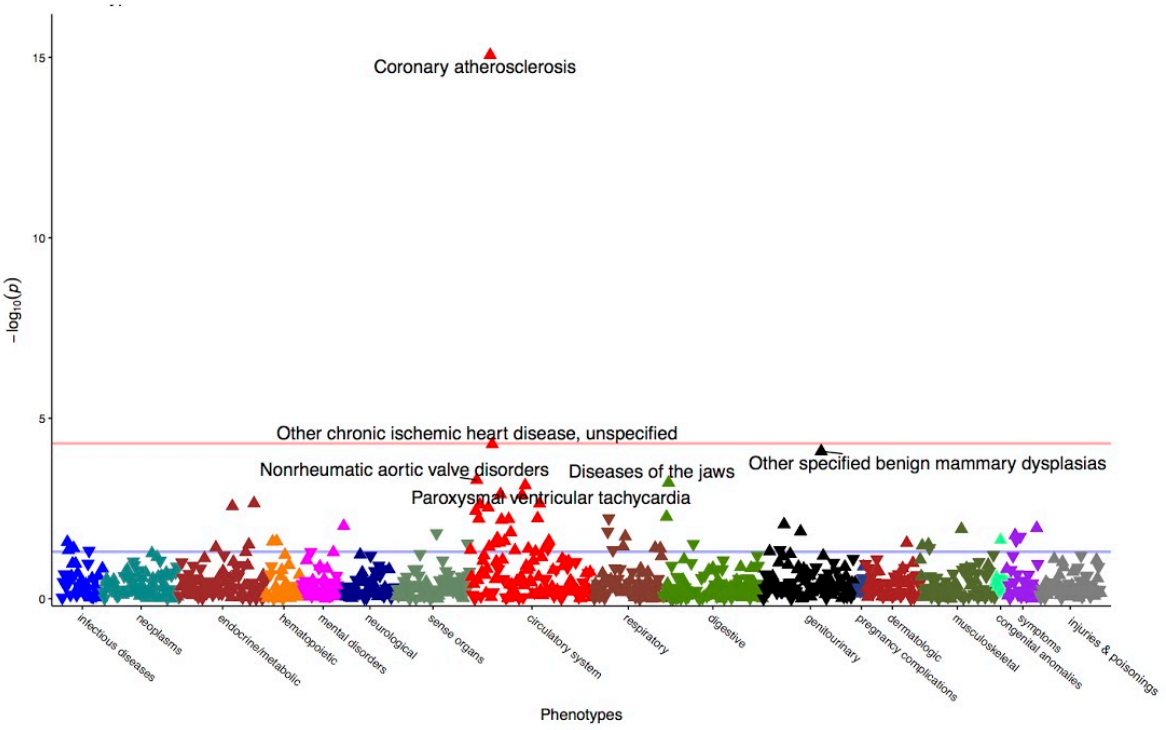


Figure 4. Comparative PheWAS of lipoprotein(a) (*LPA*) genetic variant, rs10455872. "Coronary atherosclerosis" (phecode 411.4) and "Other chronic ischemic heart disease" (phecode 411.8) were top hits associated with rs10455872 in a PheWAS analysis conducted using ICD-9-CM (top) and ICD-10-CM (bottom) phecode maps. Analyses were adjusted for age, sex, and race.

Discussion

Main Findings: Maps of ICD-10 and ICD-10-CM Codes to Phecodes have High Coverage and Yield Similar Results as the ICD-9-CM Phecode Map.
In this study, we described the process of mapping ICD-10 and ICD-10-CM codes to phecodes, and evaluated the results of the new maps in two databases. These results show that the majority of the ICD-10 and ICD-10-CM codes used in EHRs were mapped to phecodes. Our analyses suggest that researchers can expect that phenotypes sourced using the ICD-10-CM phecode map will be similar to those sourced from the gold-standard ICD-9-CM phecode map. As the use of ICD-10 and ICD-10-CM codes increases, so does the need for convenient and reliable methods of aggregating codes to represent clinically meaningful phenotypes.

Since the introduction of phecodes, many studies have demonstrated the value of aggregating ICD-9-CM codes for genetic association studies. These maps will allow biomedical researchers to leverage clinical data represented by ICD-10 and ICD-10-CM codes for their large-scale PheWAS in the EHR. They will also allow researchers to combine phenotypes as phecodes mapped from ICD-9 and ICD-10 based coding systems, thereby increasing the size of their patient cohorts and statistical power of their studies. The maps are available from https://phewascatalog.org [33] and incorporated in the PheWAS R package, version 0.99.5 (https://github.com/PheWAS/PheWAS) [13,34].

ICD-10 and ICD-10-CM Codes not Mapped to Phecodes
Analysis of the unmapped ICD-10 codes demonstrates a possible area of expansion for phecodes. The ICD-10 phecode map did not include medical concepts representing personal history or family history of disease.

We observed that a majority of the unmapped ICD-10-CM codes represented concepts that we did not expect to have phecode equivalents. Majority of the codes were from ICD-10-CM chapters 20 "External causes of morbidity" and 21 "Factors influencing health status and contact with health services". Codes from chapter 19 "Injury, poisoning, and certain other consequences of external causes" also made up a large proportion of unmapped codes, such as ICD-10-CM T38.3X6A "Underdosing of insulin and oral hypoglycemic [antidiabetic] drugs, initial encounter". We did not expect ICD-10-CM T38.3X6A to map to a phecode, as it is an encounter code that is not relevant to PheWAS. Three-digit codes that are not frequently used for reimbursement purposes, such as ICD-10-CM I67 "Other cerebrovascular diseases", also made up a large number of unmapped codes. A few potential clinically meaningful phenotypes, such as

ICD-10-CM O04.6 “Delayed or excessive hemorrhage following [induced] termination of pregnancy”, were unmapped and represent areas of potential expansion for phecodes.

ICD-10-CM Phecode Map Phenotype Reproducibility Analysis
In general, our analysis suggests that in the majority of the cases in which phenotypes are not reproduced in the ICD-10-CM observation period are not due to errors in the ICD-10-CM phecode map. This study’s reproducibility analysis (Table 2) demonstrates that the vast majority of patients (70-75%) with phecodes of four chronic diseases sourced from ICD-9-CM codes were also phenotype cases in the ICD-10-CM era. In comparison, when the same experiment is repeated for an acute disease (Intestinal infection), a minority (<10%) of patients had the same phenotype in the ICD-10-CM period.

Using the ICD-9-CM and ICD-10-CM maps, PheWAS found significant genetic associations with similar effect sizes for coronary atherosclerosis and chronic ischemic heart disease (Figure 4). Results of this analysis provide additional support for the accuracy of the ICD-10-CM map when compared to the gold-standard ICD-9-CM phecode map.

PheWAS Using ICD-10 Phecode Map
Two published studies have used the ICD-10 phecode map to identify genotype-phenotype associations using UKBB data. Zhou et al. used the map to demonstrate a method that adjusts for case-control imbalances in a large genome-wide PheWAS [35]. Li et al. used the same map to estimate the causal effects of elevated serum uric acid across the phenome [12].

Utilization of Phecodes Outside of PheWAS
In addition to being employed for PheWAS, phecodes have been used to answer a range of questions in biomedicine. Phecodes have been used to identify features in radiographic images that are associated with disease phenotypes [36], and used in machine learning models to improve cardiovascular disease prediction [37]. In a recent study to understand public opinion about diseases, Huang et al. identified articles about diseases and mapped them to phecodes [38]. Motivated by the difficulties in automatically translating diagnosis codes in the EHR, Shi et al. used phecodes to map ICD-9-CM diagnosis codes from one health system to another [39]. Phecodes have also been applied to identify conditions for aggregation in “phenotype risk scores”, much as SNPs are aggregated as a genetic risk score, to identify Mendelian diseases and determine pathogenicity of genetic variants [40].

Related Work
The Clinical Classification Software (CCS) is another maintained system for aggregating ICD codes into clinically meaningful phenotypes. CCS was originally developed by the Agency for Healthcare Research and Quality (AHRQ) to cluster ICD-9-CM diagnosis and procedure codes to a smaller number of clinically meaningful categories [41]. CCS has been used for many purposes, such as to measure outcomes [42] and to predict future health care usage [43]. In a previous study, we showed that phecodes better

aligned with diseases mentioned in clinical practice and that are relevant to genomic studies, than CCS for ICD-9-CM (CCS9) codes [20]. We found that phecodes outperform CCS9 codes, in part because CCS9 was not as granular as phecodes. Since CCS for ICD-10-CM (CCS10) is of similar granularity as CCS9 (283 vs. 285 disease groups) [44], we believe that the phecode map would likely still better represent clinically meaningful phenotypes in genetic research.

Limitations

This study has limitations. First, only 84.14% (1570/1866) of phecodes are mapped to at least one ICD-10 code. This may be due in part to the automated strategy that we used to map ICD-10 to ICD-9-CM. Second, the VUMC data are from a single site, thereby making it difficult to generalize the results of our accuracy studies (e.g. phenotype reproducibility analysis and *LPA* SNP PheWAS) to patient cohorts in other EHRs. Third, we have not yet manually reviewed all of the mappings in these "beta" phecode maps, and our assumptions that the manually-reviewed resources (e.g. NLM and OHDSI) are highly accurate could have affected the accuracy of the new phecode maps. For example, in the 2009 ICD-10-CM to ICD-9-CM GEMS, >90% of the mappings were "approximate" (i.e. non-equivalent) [15]. For this study's purposes, we aimed to maximize phecode coverage of ICD source codes, and thus included both equivalent and non-equivalent 2018 GEMS translations, which could have decreased mapping performance.

Fourth, our automated approach to map >80,000 ICD-10-CM and >9,000 ICD-10 codes to phecodes with minimal human-engineering could have decreased the accuracy of the final maps. Hripcsak et al. [45] recently evaluated the effects of translating ICD-9-CM codes to SNOMED CT codes on the creation of patient cohorts. In general, they found that mapping source billing codes to a standard clinical vocabulary (e.g. ICD-9-CM to SNOMED CT) did not greatly affect cohort selection. Their findings suggested that optimized domain knowledge-engineered mappings outperformed simple automated translations between clinical vocabularies. Using four phenotype concept sets, they showed that automated mappings resulted in errors of up to 10% and that domain-knowledge engineered mappings to have errors of <.5%. Other studies have also found that mapping performance is generally better with smaller value sets [17]. To create a more comprehensive and accurate map between ICD-9-CM and ICD-10-CM, future mapping studies could consider using an iterative forward and backward mapping approach using GEMS [17].

Future Directions

Currently, if an ICD-10 or ICD-10-CM code maps to ≥2 codes unlinked phecodes, we keep all of the mappings. In subsequent studies, it will be important to further scrutinize these mappings to ensure accuracy through manual review. As new ICD-10-CM codes are released, we plan to assess their relevance to clinical practice and genetic research, and decide whether we should translate them to phecodes. We intend to address the unmapped source codes (e.g. ICD-10-CM E78.41 "Elevated Lipoprotein(a)") by potentially expanding the phecode system, and to systematically evaluate the mappings with input from users.

Conclusions

In this paper, we introduced our work on mapping ICD-10 and ICD-10-CM codes to phecodes. We provide initial “beta” maps with high coverage of EHR data in two large databases. Results from this study suggested that the ICD-10-CM phecode map created phenotypes similar to those generated by the ICD-9-CM phecode map. These mappings will enable researchers to leverage accumulated ICD-10 and ICD-10-CM data in the EHR for large PheWAS.

Acknowledgements

P.W., A.G., J.C.D., and W.Q.W. contributed to the design of the studies. P.W., A.G., X.M., X.L., H.C., E.T, T.V., J.Z., J.C.D., and W.Q.W. analyzed the data. P.W. and A.G. were responsible for the literature review. A.G., X.M., X.L., and E.T. retrieved the raw data. P.W., A.G., R.C., L.B., J.C.D., E.T., and W.Q.W interpreted the data. P.W., A.G., J.C.D., and W.Q.W drafted the initial manuscript. P.W., A.G., J.C.D., and W.Q.W were involved in the creation and design of figures and tables. All authors revised the document and gave final approval for publication. The project was supported by NIH grant R01 LM 010685, R01 HL133786, T32 GM007347, T15 LM007450, P50 GM115305, and AHA Scientist Development Grant 16SDG27490014. The dataset used in the analyses described were obtained from Vanderbilt University Medical Center’s BioVU, which is supported by institutional funding and by the Vanderbilt CTSA grant ULTR000445 from NCATS/NIH. This research was also conducted using the UK Biobank Resource under Application Number 10775. The work conducted in Edinburgh was supported by funding for the infrastructure and staffing of the Edinburgh CRUK Cancer Research Centre. E.T. is supported by a CRUK Career Development Fellowship (C31250/ A22804). X.M. and X.L. are supported by the China Scholarship Council Studentships. We thank those individuals who manually reviewed the various maps that we used in this study [18,22,24–26]. We also thank the peer-reviewers who provided feedback for this manuscript.

Conflicts of Interest

None declared

Abbreviations

PheWAS: phenome-wide association studies
EHR: electronic health record
ICD: International Classification of Diseases
AHRQ: Agency for Healthcare Research and Quality
CM: Clinical Modification
WHO: World Health Organization
NCHS: National Center for Health Statistics
UMLS: Unified Medical Language System
GEMS: General Equivalence Mappings
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
CUI: Concept Unique Identifier
OHDSI: Observational Health Data Sciences and Informatics

NLM: National Library of Medicine
VUMC: Vanderbilt University Medical Center
UKBB: UK Biobank
SD: standard deviation
OR: odds ratio
LPA: lipoprotein(a)
SNP: single nucleotide polymorphism
CCS: Clinical Classification Software

Multimedia Appendix 1: ICD-10-CM reproducibility analysis, manual chart review results

Phenotype	Group	Number of People
Hypertension	absence of ICD-10-CM only	2
Hyperlipidemia	absence of ICD-10-CM only	4
Type 2 diabetes	absence of ICD-10-CM only	4
Hypertension	short observation window	1
Hyperlipidemia	short observation window	4
Type 2 diabetes	short observation window	3
Hypertension	limited number of visits/specialist	7
Hyperlipidemia	limited number of visits/specialist	2
Type 2 diabetes	limited number of visits/specialist	1
Hypertension	inconsistent diagnosis	0
Hyperlipidemia	inconsistent diagnosis	0
Type 2 diabetes	inconsistent diagnosis	2

References

1. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, Pacheco JA, Tromp G, Pathak J, Carrell DS, Ellis SB, Lingren T, Thompson WK, Savova G, Haines J, Roden DM, Harris PA, Denny JC. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc [Internet] 2016 Nov;23(6):1046–1052. PMID:27026615

2. Gamazon ER, Segrè AV, van de Bunt M, Wen X, Xi HS, Hormozdiari F, Ongen H, Konkashbaev A, Derks EM, Aguet F, Quan J, GTEx Consortium, Nicolae DL, Eskin E, Kellis M, Getz G, McCarthy MI, Dermitzakis ET, Cox NJ, Ardlie KG. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat Genet [Internet] 2018 Jul;50(7):956–967. PMID:29955180

3. Nielsen JB, Thorolfsdottir RB, Fritsche LG, Zhou W, Skov MW, Graham SE, Herron TJ, McCarthy S, Schmidt EM, Sveinbjornsson G, Surakka I, Mathis MR, Yamazaki M, Crawford RD, Gabrielsen ME, Skogholt AH, Holmen OL, Lin M, Wolford BN, Dey R, Dalen H, Sulem P, Chung JH, Backman JD, Arnar DO, Thorsteinsdottir U, Baras A, O'Dushlaine C, Holst AG, Wen X, Hornsby W, Dewey FE, Boehnke M, Kheterpal S, Mukherjee B, Lee S, Kang HM, Holm H, Kitzman J, Shavit JA, Jalife J, Brummett CM, Teslovich TM, Carey DJ, Gudbjartsson DF, Stefansson K, Abecasis GR, Hveem K, Willer CJ. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. Nat Genet [Internet] 2018 Sep;50(9):1234–1239. PMID:30061737

4. Simonti CN, Vernot B, Bastarache L, Bottinger E, Carrell DS, Chisholm RL, Crosslin DR, Hebbring SJ, Jarvik GP, Kullo IJ, Li R, Pathak J, Ritchie MD, Roden DM, Verma SS, Tromp G, Prato JD, Bush WS, Akey JM, Denny JC, Capra JA. The phenotypic legacy of admixture between modern humans and Neandertals. Science [Internet] 2016 Feb 12;351(6274):737–741. PMID:26912863

5. Diogo D, Bastarache L, Liao KP, Graham RR, Fulton RS, Greenberg JD, Eyre S, Bowes J, Cui J, Lee A, Pappas DA, Kremer JM, Barton A, Coenen MJH, Franke B, Kiemeney LA, Mariette X, Richard-Miceli C, Canhão H, Fonseca JE, de Vries N, Tak PP, Crusius JBA, Nurmohamed MT, Kurreeman F, Mikuls TR, Okada Y, Stahl EA, Larson DE, Deluca TL, O'Laughlin M, Fronick CC, Fulton LL, Kosoy R, Ransom M, Bhangale TR, Ortmann W, Cagan A, Gainer V, Karlson EW, Kohane I, Murphy SN, Martin J, Zhernakova A, Klareskog L, Padyukov L, Worthington J, Mardis ER, Seldin MF, Gregersen PK, Behrens T, Raychaudhuri S, Denny JC, Plenge RM. TYK2 protein-coding variants protect against rheumatoid arthritis and autoimmunity, with no evidence of major pleiotropic effects on non-autoimmune complex traits. PLoS One [Internet] 2015 Apr 7;10(4):e0122271. PMID:25849893

6. Rastegar-Mojarad M, Ye Z, Kolesar JM, Hebbring SJ, Lin SM. Opportunities for drug repositioning from phenome-wide association studies. Nat Biotechnol [Internet] 2015 Apr;33(4):342–345. PMID:25850054

7. Millard LAC, Davies NM, Timpson NJ, Tilling K, Flach PA, Davey Smith G. MR-PheWAS: hypothesis prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization. Sci Rep [Internet] 2015 Nov 16;5:16645. PMID:26568383

8. Ehm MG, Aponte JL, Chiano MN, Yerges-Armstrong LM, Johnson T, Barker JN, Cook SF, Gupta A, Hinds DA, Li L, Nelson MR, Simpson MA, Tian C, McCarthy LC, Rajpal DK, Waterworth DM. Phenome-wide association study using research participants' self-reported data provides insight into the Th17 and IL-17 pathway. PLoS One [Internet] 2017 Nov 1;12(11):e0186405. PMID:29091937

9. Liu J, Ye Z, Mayer JG, Hoch BA, Green C, Rolak L, Cold C, Khor S-S, Zheng X, Miyagawa T, Tokunaga K, Brilliant MH, Hebbring SJ. Phenome-wide association study maps new diseases to the human major histocompatibility complex region. J Med Genet [Internet] 2016 Oct;53(10):681–689. PMID:27287392

10. Neuraz A, Chouchana L, Malamut G, Le Beller C, Roche D, Beaune P, Degoulet P, Burgun A, Loriot M-A, Avillach P. Phenome-wide association studies on a quantitative trait: application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. PLoS Comput Biol [Internet] 2013 Dec 26;9(12):e1003405. PMID:24385893

11. Doshi-Velez F, Ge Y, Kohane I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. Pediatrics [Internet] 2014 Jan;133(1):e54–63. PMID:24323995

12. Li X, Meng X, Spiliopoulou A, Timofeeva M, Wei W-Q, Gifford A, Shen X, He Y, Varley T, McKeigue P, Tzoulaki I, Wright AF, Joshi P, Denny JC, Campbell H, Theodoratou E. MR-PheWAS: exploring the causal effect of SUA level on multiple disease outcomes by using genetic instruments in UK Biobank. Ann Rheum Dis [Internet] 2018 Jul;77(7):1039–1047. PMID:29437585

13. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics [Internet] 2014 Aug 15;30(16):2375–2376. PMID:24733291

14. Phecode Map 1.2 with ICD-9 Codes [Internet]. PheWAS Catalog. [cited 2019 Jul 17]. Available from: https://phewascatalog.org/phecodes

15. Steindel SJ. International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. J Am Med Inform Assoc [Internet] 2010 May;17(3):274–282. PMID:20442144

16. Topaz M, Shafran-Topaz L, Bowles KH. ICD-9 to ICD-10: evolution, revolution, and current debates in the United States. Perspect Health Inf Manag [Internet] 2013 Apr 1;10(Spring):1d. PMID:23805064

17. Fung KW, Richesson R, Smerek M, Pereira KC, Green BB, Patkar A, Clowse M, Bauck A, Bodenreider O. Preparing for the ICD-10-CM Transition: Automated Methods for Translating ICD Codes in Clinical Phenotype Definitions. EGEMS (Wash DC) [Internet] 2016 Apr 12;4(1):1211. PMID:27195309

18. Wilder V. UMLS 2018AA Release Available [Internet]. U.S. National Library of Medicine; 2018 [cited 2019 Jul 17]. Available from: https://www.nlm.nih.gov/pubs/techbull/mj18/mj18_umls_2018aa_release.html

19. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, Field JR, Pulley JM, Ramirez AH, Bowton E, Basford MA, Carrell DS, Peissig PL, Kho AN, Pacheco JA, Rasmussen LV, Crosslin DR, Crane PK, Pathak J, Bielinski SJ, Pendergrass SA, Xu H, Hindorff LA, Li R, Manolio TA, Chute CG, Chisholm RL, Larson EB, Jarvik GP, Brilliant

MH, McCarty CA, Kullo IJ, Haines JL, Crawford DC, Masys DR, Roden DM. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol [Internet] 2013 Dec;31(12):1102–1110. PMID:24270849

20. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, Cox NJ, Roden DM, Denny JC. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLoS One [Internet] 2017 Jul 7;12(7):e0175508. PMID:28686612

21. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med [Internet] 2015 Mar;12(3):e1001779. PMID:25826379

22. 2018-ICD-10-CM-and-GEMs [Internet]. 2017 [cited 2019 Jul 5]. Available from: https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-CM-and-GEMs.html

23. Fung KW, Bodenreider O. Utilizing the UMLS for semantic mapping between terminologies. AMIA Annu Symp Proc [Internet] 2005;266–270. PMID:16779043

24. ICD9CM Observational Health Data Sciences and Informatics [Internet]. Observational Health Data Sciences and Informatics. [cited 2019 Jul 17]. Available from: https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:icd9cm

25. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong ICK, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li Y-C, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform [Internet] 2015;216:574–578. PMID:26262116

26. SNOMED Terminology Solutions. SNOMED CT to ICD-9-CM Rule Based Mapping to Support Reimbursement [Internet]. Unified Medical Language System. [cited 2019 Apr 7]. Available from: https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd9cm_reimburse.html

27. Giuse DA. Supporting communication in an integrated patient record system. AMIA Annu Symp Proc [Internet] 2003;1065. PMID:14728568

28. Nordestgaard BG, Chapman MJ, Ray K, Borén J, Andreotti F, Watts GF, Ginsberg H, Amarenco P, Catapano A, Descamps OS, Fisher E, Kovanen PT, Kuivenhoven JA, Lesnik P, Masana L, Reiner Z, Taskinen M-R, Tokgözoglu L, Tybjærg-Hansen A, European Atherosclerosis Society Consensus Panel. Lipoprotein(a) as a cardiovascular risk factor: current status. Eur Heart J [Internet] 2010 Dec;31(23):2844–2853. PMID:20965889

29. Zhao J, Feng Q, Wu P, Warner JL, Denny JC, Wei W-Q. Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of Lipoprotein(a) (LPA). PLoS One [Internet] 2019 Feb 13;14(2):e0212112. PMID:30759150

30. Wei W-Q, Li X, Feng Q, Kubo M, Kullo IJ, Peissig PL, Karlson EW, Jarvik GP, Lee MTM, Shang N, Larson EA, Edwards T, Shaffer CM, Mosley JD, Maeda S, Horikoshi M, Ritchie M, Williams MS, Larson EB, Crosslin DR, Bland ST, Pacheco JA, Rasmussen-Torvik LJ, Cronkite D, Hripcsak G, Cox NJ, Wilke RA, Stein CM, Rotter JI, Momozawa Y, Roden DM, Krauss RM, Denny JC. LPA Variants Are Associated With Residual Cardiovascular Risk in Patients Receiving Statins. Circulation [Internet] 2018 Oct 23;138(17):1839–1849. PMID:29703846

31. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther [Internet] 2008 Sep;84(3):362–369. PMID:18500243

32. Wilder V. UMLS 2018AA [Internet]. NLM Technical Bulletin. [cited 2019 Apr 7]. Available from: https://www.nlm.nih.gov/pubs/techbull/mj18/mj18_umls_2018aa_release.html

33. PheWAS - Phenome Wide Association Studies [Internet]. [cited 2019 Aug 2]. Available from: https://phewascatalog.org/

34. PheWAS R Package, GitHub Repository [Internet]. GitHub. Available from: https://github.com/PheWAS/PheWAS

35. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, Bastarache LA, Wei W-Q, Denny JC, Lin M, Hveem K, Kang HM, Abecasis GR, Willer CJ, Lee S. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet [Internet] 2018 Sep;50(9):1335–1341. PMID:30104761

36. Chaganti S, Mawn LA, Kang H, Egan J, Resnick SM, Beason-Held LL, Landman BA, Lasko T. Electronic Medical Record Context Signatures Improve Diagnostic Classification using Medical Image Computing. IEEE J Biomed Health Inform [Internet] 2018 Dec 28; PMID:30602428

37. Zhao J, Feng Q, Wu P, Lupu RA, Wilke RA, Wells QS, Denny JC, Wei W-Q. Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. Sci Rep [Internet] 2019 Jan 24;9(1):717. PMID:30679510

38. Huang M, ElTayeby O, Zolnoori M, Yao L. Public Opinions Toward Diseases: Infodemiological Study on News Media Data. J Med Internet Res [Internet] 2018 May 8;20(5):e10047. PMID:29739741

39. Shi X, Li X, Cai T. Spherical Regression under Mismatch Corruption with Application to Automated Knowledge Translation [Internet]. arXiv [statME]. 2018. Available from: http://arxiv.org/abs/1810.05679

40. Bastarache L, Hughey JJ, Hebbring S, Marlo J, Zhao W, Ho WT, Van Driest SL, McGregor TL, Mosley JD, Wells QS, Temple M, Ramirez AH, Carroll R, Osterman T, Edwards T, Ruderfer D, Velez Edwards DR, Hamid R, Cogan J, Glazer A, Wei W-Q, Feng Q, Brilliant M, Zhao ZJ, Cox NJ, Roden DM, Denny JC. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science [Internet] 2018 Mar 16;359(6381):1233–1239. PMID:29590070

41. HCUP-US Tools & Software Page [Internet]. [cited 2019 Jul 15]. Available from: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp

42. Sabbatini AK, Kocher KE, Basu A, Hsia RY. In-Hospital Outcomes and Costs Among Patients Hospitalized During a Return Visit to the Emergency Department. JAMA [Internet] 2016 Feb 16;315(7):663–671. PMID:26881369

43. Hu Z, Hao S, Jin B, Shin AY, Zhu C, Huang M, Wang Y, Zheng L, Dai D, Culver DS, Alfreds ST, Rogow T, Stearns F, Sylvester KG, Widen E, Ling X. Online Prediction of Health Care Utilization in the Next Six Months Based on Electronic Health Record Information: A Cohort and Validation Study. J Med Internet Res [Internet] 2015 Sep 22;17(9):e219. PMID:26395541

44. Beta Clinical Classifications Software (CCS) for ICD-10-CM/PCS [Internet]. [cited 2019 Jul 5]. Available from: https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp

45. Hripcsak G, Levine ME, Shang N, Ryan PB. Effect of vocabulary mapping for conditions on phenotype cohorts. J Am Med Inform Assoc [Internet] 2018 Dec 1;25(12):1618–1625. PMID:30395248